Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Semestre Primavera 2006
Cálculo Numérico
MA33A-01

Profesor: Gonzalo Hernández.
Auxiliar: Gonzalo Ríos.
Fecha: 24 de Agosto

Pauta Control 1

1) Representación Numerica

(a) Considere una representación floating point tipo inicial, donde el número real más cercano a cero es 2^{-23}. Se sabe que la cantidad de bits de la mantisa es el doble que el número de bits usados en el exponente.

i) Explique cuantos bits son necesarios para implementar esta codificación y como se distribuyen.

ii) Sea $x = a_1 a_2 ... a_n$ la representación floating point de un número real, donde n es el números de bits utilizados por la codificación y:

$$a_1 = 1, \quad a_k = (a_{k-1} + k) \bmod 2, \ k = 2...n$$

Determine el real representado por x.

iii) Calcule el real máximo representable en esta codificación.

(b) Considere la codificación floating point actual de 32 bits de la forma:

$$fl(x) = (-1)^s 2^z (1 + m)$$

donde s el bit de signo, $z \in [-126, 127]$ el exponente y m la mantisa de 23 bits.

Suponga que se redondea un número $x > 0$ para llegar a su codificación floating point.

i) Demuestre que:

$$-2^{z-n} \leq x - fl(x) \leq 2^{z-n}$$

donde n es la cantidad de bits de la mantisa.

ii) Demuestre que $x \geq 2^z$ y que por lo tanto:

$$\frac{|x - fl(x)|}{x} \leq 2^{-n}$$

Respuesta

1) Representación Numérica

(a) Representación floating point tipo inicial

i) Como el número real más cercano a cero representable en una codificacion binaria siempre es $2^{-m} \times 2^{-(2^e-1)}$, donde m es la cantidad de bits de la mantisa, y e la cantidad de bits del exponente, esto equivale a $2^{-(2^e-1)-m}$, y como $m = 2e$ queda un sistema:
$2^{-(2^e-1)-m} = 2^{-23} \Longrightarrow \begin{matrix} 2^e - 1 + m = 23 \\ m = 2e \end{matrix} \Longrightarrow 2^e + 2e = 24 \Longrightarrow \begin{matrix} e = 4 \\ m = 2e = 8 \end{matrix}$
Sumando el bit para el signo de la mantisa y el bit del signo del exponente, la codificación necesita $2 + 4 + 8 = 14$ bits.

ii) En total son 14 bits, entonces

a_1 1
a_2 $(1+2) \bmod 2 = 1$
a_3 $(1+3) \bmod 2 = 0$
a_4 $(0+4) \bmod 2 = 0$
a_5 $(0+5) \bmod 2 = 1$
a_6 $(1+6) \bmod 2 = 1$
a_7 $(1+7) \bmod 2 = 0$
a_8 $(0+8) \bmod 2 = 0$
a_9 $(0+9) \bmod 2 = 1$
a_{10} $(1+10) \bmod 2 = 1$
a_{11} $(1+11) \bmod 2 = 0$
a_{12} $(0+12) \bmod 2 = 0$
a_{13} $(0+13) \bmod 2 = 1$
a_{14} $(1+14) \bmod 2 = 1$

Número real: $x = 1\ 1\ 0011\ 00110011$

Signo Mantisa: $-$

Signo Exponente: $-$

Exponente:$0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 3$

Mantisa: $0 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} + 1 \times 2^{-4} + 0 \times 2^{-5} + 0 \times 2^{-6} + 1 \times 2^{-7} + 1 \times 2^{-8} = \frac{1}{8} + \frac{1}{16} + \frac{1}{128} + \frac{1}{256} = \frac{32+16+2+1}{256} = \frac{51}{256}$

Numero: $x = -\frac{51}{256} \times 2^{-3} = -\frac{51}{256} \times \frac{1}{8} = -\frac{51}{2048} = -0.024\,902\,343\,75$

iii) El número real máximo representable en una codificación binario tipo inicial es de la forma $M = (1 - 2^{-m}) \times 2^{(2^e - 1)}$.Como $m = 8$ y $e = 4$, entonces
$M = (1 - 2^{-8}) \times 2^{(2^4 - 1)} = 2^{15} - 2^7 = 32640$

(b) Codificación floating point actual

i) Dado un real $x > 0$, ese se pasa a base binaria, con una mantisa entre 1 y 2, es decir $x = 2^z(1 + m^*)$,con $m^* \in [0, 1]$. Ahora, al aproximar queda de la forma $fl(x) = 2^z(1 + m)$ Entonces, se puede hacer una transformación de codificación tipo actual a tipo inicial tomando $x = 2^{z+1}(\frac{1+m^*}{2})$, $fl(x) = 2^{z+1}(\frac{1+m}{2})$, en donde se cumple que la nueva mantisa es mayor que $\frac{1}{2}$ y menor que 1, y ahora tiene $n + 1$ cifras significativas. Como vimos en clase, en la codificación tipo inicial se tiene que $E_{rel}(x, fl(x)) \leq \frac{b}{2} \times b^{-t}$, donde b es la base y t es la cantidad de cifras significativas de la mantisa, entondes en nuestro caso, $E_{rel}(x, fl(x)) \leq 2^{-(n+1)} \Longrightarrow \frac{|x - fl(x)|}{x} \leq 2^{-(n+1)} \Longrightarrow |x - fl(x)| \leq x2^{-(n+1)}$
$\Longrightarrow |x - fl(x)| \leq x2^{-(n+1)} = 2^{z+1}(\frac{1+m^*}{2}) \times 2^{-(n+1)} \leq 2^{z+1-n-1} = 2^{z-n}$
$\Longrightarrow -2^{z-n} \leq x - fl(x) \leq 2^{z-n}$

ii) Como $x = 2^z(1 + m^*)$, con $m^* \in [0, 1] \Longrightarrow (1 + m^*) \geq 1 \Longrightarrow x = 2^z(1 + m^*) \geq 2^z$
$\Longrightarrow \frac{1}{x} \leq \frac{1}{2^z}$
Entonces, $E_{rel}(x, fl(x)) = \frac{|x - fl(x)|}{x} \leq \frac{2^{z-n}}{x} \leq \frac{2^{z-n}}{2^z} \leq 2^{-n}$

2) Aproximación y Errores

(a) Determine la propagación de errores de las siguientes operaciones matemáticas:

i) $\phi(x, y) = 1 + \sin(x^2 + y^2), \quad x, y \in [0, 1]$

ii) $\phi(x, y, z) = \sqrt{xyz}$

Cuáles de ellas son estables/inestables ? Fundamente su respuesta.

(b) Considere las siguientes aproximaciones numéricas de la derivada de una función a variable real $f(x)$ en un punto x_0: (para $h \approx 0$)

$$\frac{df(x_0)}{dx} \approx \frac{f(x_0 + h) - f(x_0)}{h} \qquad \frac{df(x_0)}{dx} \approx \frac{f(x_0 + h) - f(x_0 - h)}{2h}$$

i) Haga un análisis de propagación de error de la primera fórmula con $\varepsilon_h = 0$.
Obs: Si aparece $\frac{f(x+h)-f(x)}{h}$, aproxímelo por $f^{'}(x)$.

ii) Para una aritmética finita de 5 cifras significativas con redondeo, aproxime la derivada de $f(x) = x \ln x$ en el punto $x_0 = 10.125$, usando la segunda fórmula con $h = 0.001$. Compare este valor aproximado con el valor exacto de la derivada. Calcule el error absoluto y relativo.

Respuesta

2) Aproximación y Errores

(a) Estabilidad: $\varepsilon_\phi = \frac{x}{\phi(x,y)}\frac{\partial\phi}{\partial x}\varepsilon_x + \frac{y}{\phi(x,y)}\frac{\partial\phi}{\partial y}\varepsilon_y$

i) $\phi(x,y) = 1 + \sin(x^2+y^2)$
$\frac{\partial\phi}{\partial x} = \cos(x^2+y^2)2x$
$\frac{\partial\phi}{\partial y} = \cos(x^2+y^2)2y$
$\varepsilon_\phi = \frac{x}{\phi(x,y)}\frac{\partial\phi}{\partial x}\varepsilon_x + \frac{y}{\phi(x,y)}\frac{\partial\phi}{\partial y}\varepsilon_y = \frac{\cos(x^2+y^2)2x^2}{1+\sin(x^2+y^2)}\varepsilon_x + \frac{\cos(x^2+y^2)2y^2}{1+\sin(x^2+y^2)}\varepsilon_y$
Como $x, y \in [0,1]$, entonces el errores se podría acotar por su mayor valor que tome en ese intervalo, siempre y cuando exista. La única forma en que se indetermina es en el caso que $1 + \sin(x^2+y^2) = 0$, que solo podría suceder si $\sin(x^2+y^2) = -1$, que sería cuando $x^2 + y^2 = \frac{3\pi}{2}$, pero como $x, y \in [0,1]$, $\max(\ x^2 + y^2) = 2 < \frac{3\pi}{2}$, lo cual es estable.r

ii) $\phi(x,y,z) = \sqrt{xyz}$; $\varepsilon_\phi = \frac{x}{\phi(x,y,z)}\frac{\partial\phi}{\partial x}\varepsilon_x + \frac{y}{\phi(x,y,z)}\frac{\partial\phi}{\partial y}\varepsilon_y + \frac{z}{\phi(x,y,z)}\frac{\partial\phi}{\partial z}\varepsilon_z$
$\frac{\partial\phi}{\partial x} = \frac{1}{2}\sqrt{\frac{yz}{x}}$
$\frac{\partial\phi}{\partial y} = \frac{1}{2}\sqrt{\frac{xz}{y}}$
$\frac{\partial\phi}{\partial z} = \frac{1}{2}\sqrt{\frac{xy}{z}}$
$\varepsilon_\phi = \frac{x}{\phi(x,y,z)}\frac{\partial\phi}{\partial x}\varepsilon_x + \frac{y}{\phi(x,y,z)}\frac{\partial\phi}{\partial y}\varepsilon_y + \frac{z}{\phi(x,y,z)}\frac{\partial\phi}{\partial z}\varepsilon_z = \frac{x}{\sqrt{xyz}}\frac{1}{2}\sqrt{\frac{yz}{x}}\varepsilon_x + \frac{y}{\sqrt{xyz}}\frac{1}{2}\sqrt{\frac{xz}{y}}\varepsilon_y + \frac{z}{\sqrt{xyz}}\frac{1}{2}\sqrt{\frac{xy}{z}}\varepsilon_z =$
$\frac{1}{2}(\varepsilon_x + \varepsilon_y + \varepsilon_z)$

(b) $\frac{df(x_0)}{dx} \simeq \frac{f(x_0+h)-f(x_0-h)}{2h}$

i) $f'(x) = \phi(x,h) = \frac{f(x+h)-f(x)}{h}$
$\phi_0(x) = f(x)$
$\varepsilon_0 = x\frac{f'(x)}{f(x)}\varepsilon_x$
$\phi_1(x,h) = f(x+h)$
$\varepsilon_1 = x\frac{f'(x+h)}{f(x+h)}\varepsilon_x + h\frac{f'(x+h)}{f(x+h)}\varepsilon_h = x\frac{f'(x+h)}{f(x+h)}\varepsilon_x$
$\phi_2(f(x+h), f(x)) = f(x+h) - f(x)$
$\varepsilon_2 = \frac{f(x+h)}{f(x+h)-f(x)}\varepsilon_1 - \frac{f(x)}{f(x+h)-f(x)}\varepsilon_0 = \frac{f(x+h)}{f(x+h)-f(x)}x\frac{f'(x+h)}{f(x+h)}\varepsilon_x - \frac{f(x)}{f(x+h)-f(x)}x\frac{f'(x)}{f(x)}\varepsilon_x$
$= \frac{xf'(x+h)\varepsilon_x}{f(x+h)-f(x)} - \frac{xf'(x)\varepsilon_x}{f(x+h)-f(x)} = \frac{f'(x+h)-f'(x)}{f(x+h)-f(x)}x\varepsilon_x$
$f'(x,h) = \frac{\phi_2}{2h}$
$\varepsilon_{f'} = \varepsilon_2 - \varepsilon_h = \varepsilon_2 = \frac{f'(x+h)-f'(x)}{f(x+h)-f(x)}x\varepsilon_x = \frac{\frac{f'(x+h)-f'(x)}{h}}{\frac{f(x+h)-f(x)}{h}}x\varepsilon_x = \frac{f''(x)}{f'(x)}x\varepsilon_x$
Obs: Si se toma $\phi(x,h) = \frac{f(x+h)-f(x)}{h}$, entonces $\varepsilon_\phi = \frac{x}{\frac{f(x+h)-f(x)}{h}}\frac{f'(x+h)-f'(x)}{h}\varepsilon_x =$
$\frac{x}{f'(x)}f''(x)\varepsilon_x = \frac{f''(x)}{f'(x)}x\varepsilon_x$,
llegando al mismo resultado, pero la idea era hacerlo dividiendo por operaciones.

ii) $f(x) = x\ln x$, $x_0 = 10.125$, $h = 0.001$
$\frac{df(x_0)}{dx} \simeq \frac{f(x_0+h)-f(x_0-h)}{2h}$

A. $fl(x_0) = 10.125$
$fl(h) = 0.001$
$fl(x_0 + h) = fl(10.126) = 10.126$
$fl(\ln(10.126)) = fl(2.315\,106\,373\,547\,717\,429\,6) = 2.315\,1$
$fl(10.126 \times 2.315\,1) = fl(23.442\,702\,6) = 23.443$
$fl(x_0 - h) = fl(10.124) = 10.124$
$fl(\ln(10.124)) = fl(2.314\,908\,842\,682\,877\,619\,8) = 2.314\,9$
$fl(10.124 \times 2.314\,9) = fl(23.436\,047\,6) = 23.436$
$fl(23.443 - 23.436) = fl(0.007) = 0.007$

$fl(2 \times 0.001) = fl(0.002) = 0.002$
$fl(\frac{0.007}{0.002}) = fl(3.5) = 3.5$

B. $f'(x) = \ln x + 1$
$f'(10.125) = \ln 10.125 + 1 = 3.315\,007\,612\,992\,602\,837\,3$
$\Longrightarrow fl(f'(10.125)) = 3.315$
$E_{abs} = |3.315 - 3.5| = 0.185$
$E_{rel} = \frac{|3.315 - 3.5|}{|3.315|} = 0.055\,806\,938\,159\,879\,336\,35 < 5 \times 10^{-1}$
$\Longrightarrow$ tiene 1 cifra significativa

3) Sistemas de Ecuaciones Lineales

Dada una matriz tridiagonal, su factorización $A = LU$ puede ser llevada a la forma:

$$\begin{bmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn-1} & a_{nn} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & l_{nn-1} & l_{nn} \end{bmatrix} \begin{bmatrix} 1 & u_{12} & 0 & \cdots & 0 \\ 0 & 1 & u_{23} & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & u_{n-1n} \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

mediante el Método de Crout:

Paso 1:	Paso 2: Para $i = 2, ..., n-1$	Paso 3:
$l_{11} = a_{11}$ $u_{12} = \frac{a_{12}}{l_{11}}$	$l_{i(i-1)} = a_{i(i-1)}$ $l_{ii} = a_{ii} - l_{i(i-1)}u_{(i-1)i}$ $u_{i(i+1)} = \frac{a_{i(i+1)}}{l_{ii}}$	$l_{n(n-1)} = a_{n(n-1)}$ $l_{nn} = a_{nn} - l_{n(n-1)}u_{(n-1)n}$

(a) Calcule la factorización de Crout de la siguiente matriz:

$$A = \begin{bmatrix} 5 & -2 & 0 & 0 \\ -2 & 5 & -2 & 0 \\ 0 & -2 & 5 & -2 \\ 0 & 0 & -2 & 5 \end{bmatrix}$$

(b) Calcule la cantidad de *ops* que este método realiza aplicado a una matriz cuadrada tridiagonal de tamaño n.

Respuesta

3) Sistemas de Ecuaciones Lineales

(a) $A = \begin{bmatrix} 5 & -2 & 0 & 0 \\ -2 & 5 & -2 & 0 \\ 0 & -2 & 5 & -2 \\ 0 & 0 & -2 & 5 \end{bmatrix}$

Paso 1: $l_{11} = a_{11} = 5$
$u_{12} = \frac{a_{12}}{l_{11}} = \frac{-2}{5}$

Paso 2: $l_{21} = a_{21} = -2$
$l_{22} = a_{22} - l_{21}u_{12} = 5 - (-2)\left(\frac{-2}{5}\right) = \frac{21}{5}$
$u_{23} = \frac{a_{23}}{l_{22}} = \frac{-2}{\frac{21}{5}} = -\frac{10}{21}$
$l_{32} = a_{32} = -2$
$l_{33} = a_{33} - l_{32}u_{23} = 5 - (-2)\left(-\frac{10}{21}\right) = \frac{85}{21}$
$u_{34} = \frac{a_{34}}{l_{33}} = \frac{-2}{\frac{85}{21}} = -\frac{42}{85}$

Paso 3: $l_{43} = a_{43} = -2$
$l_{44} = a_{44} - l_{43}u_{34} = 5 - (-2)\left(-\frac{42}{85}\right) = \frac{341}{85}$

Finalmente:

$$L = \begin{bmatrix} 5 & 0 & 0 & 0 \\ -2 & \frac{21}{5} & 0 & 0 \\ 0 & -2 & \frac{85}{21} & 0 \\ 0 & 0 & -2 & \frac{341}{85} \end{bmatrix}; \; U = \begin{bmatrix} 1 & \frac{-2}{5} & 0 & 0 \\ 0 & 1 & -\frac{10}{21} & 0 \\ 0 & 0 & 1 & -\frac{42}{85} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

(b) $ops = 1 + \sum_{i=2}^{n-1} 3 \; + \; 2 = 3(n-2) + 3 = 3(n-1) \Longrightarrow O(n)$ lineal